All are Worth Words: A ViT Backbone for Diffusion Models

Year: 2022 | Citations: 481 | Authors: Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li

Abstract

Vision transformers (ViT) have shown promise in various vision tasks while the U-Net based on a convolutional neural network (CNN) remains dominant in diffusion models. We design a simple and general ViT-based architecture (named U-ViT) for image generation with diffusion models. U-ViT is characterized by treating all inputs including the time, condition and noisy image patches as tokens and employing long skip connections between shallow and deep layers. We evaluate U-ViT in unconditional and classconditional image generation, as well as text-to-image generation tasks, where U-ViT is comparable if not superior to a CNN-based U-Net of a similar size. In particular, latent diffusion models with U-ViT achieve record-breaking FID scores of 2.29 in class-conditional image generation on ImageNet 256×256, and 5.48 in text-to-image generation on MS-COCO, among methods without accessing large external datasets during the training of generative models. Our results suggest that, for diffusion-based image modeling, the long skip connection is crucial while the down-sampling and upsampling operators in CNN-based U-Net are not always necessary. We believe that U-ViT can provide insights for future research on backbones in diffusion models and benefit generative modeling on large scale cross-modality datasets.